UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

May 22, 2012

DSSD CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-07

MEMORANDUM FOR David C. Whitford
Chief, Decennial Statistical Studies Division

From: Patrick J. Cantwell *(Signed)*
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Roger Shores
James Mulligan
Robert Sands
Decennial Statistical Studies Division

Subject: 2010 Census Coverage Measurement Estimation Report:
Characteristic Imputation Results

This report is one of twelve documents providing estimation results from the 2010 Census Coverage Measurement program. This report focuses on the results of characteristic imputation for the United States. The characteristics are relationship, age, sex, race, Hispanic origin, and tenure.

For more information, contact Roger Shores on (301) 763-9282, James Mulligan on (301) 763-1978, or Robert Sands on (301) 763-4255.

cc:
DSSD CCM Contacts List

Census Coverage Measurement Estimation Report

Characteristic Imputation Results

Prepared by
Roger Shores
James Mulligan
Robert Sands

Decennial Statistical Studies Division

Table of Contents

Executive Summary

Characteristic imputation in the 2010 Census Coverage Measurement program imputed values when missing values occurred for relationship, race, Hispanic origin, age, sex, and tenure. The key distinguishing feature of characteristic imputation in the Census Coverage Measurement program is that it used the same characteristic imputation system that was applied to the 2010 Census.

The census characteristic imputation system included many rules for editing the data. No previous census coverage survey had used this editing process. The extent of editing in the P sample in the Census Coverage Measurement program ranged from none for sex and tenure to 1.3% for relationship.

A procedure that was new for the Census Coverage Measurement program was the autocoding of write-in responses for race and Hispanic origin. There were 14.7% and 4.1% of P-sample records with a write-in response for race and Hispanic origin, respectively. Race changed for 1.6%, and Hispanic origin 0.2%, of P-sample records due to the write-in responses. Most changes for race resulted from the write-in response for the Some Other Race racial category.

The highest rate of imputation in the P sample was 6.1% for age, and the lowest, 1.3% for sex. Ten percent of P-sample records had at least one characteristic imputed. The percentage of cases imputed in the Census Coverage Measurement program was higher for most categories than for their counterparts in the 2000 Accuracy and Coverage Evaluation Revision II program.

1. Introduction

This document gives the results of characteristic imputation in the 2010 Census Coverage Measurement (CCM) program for the United States. It also compares results from the CCM with those from the 2000 Accuracy and Coverage Evaluation (A.C.E.) Revision II program.

Characteristic imputation is the process by which certain missing person-level or household items are filled in for the census or the CCM. The characteristics that are subject to imputation are relationship, age, sex, race, Hispanic origin, and tenure.

The CCM consisted of the P sample and the E sample. The P sample was derived from an independent listing of housing units in the United States. The E sample consisted of a separate sample taken from census records in the same sample block clusters from which the P sample was drawn. For the P sample, the CCM used the same characteristic imputation system that was applied to the 2010 Census. This use of the same characteristic imputation system for both the census and coverage survey had not been done before. The 2000 A.C.E. Revision II, for example, used an entirely independent characteristic imputation system for the P sample. Imputation methods in the CCM E sample, however, were the same as those used in the census, which was also true for the A.C.E. Revision II.

An important feature of the census system that was new to a census coverage survey was the editing of reported values to achieve household consistency. Previous census coverage surveys did not draw on census editing rules to change reported values. In the CCM, values for characteristics were subject to change through application of the census edit rules.

The census characteristic imputation system had certain other elements that had not been a part of missing data procedures for previous census coverage surveys. One example was the use of first name and surname to aid in the imputation of sex and Hispanic origin, respectively. Another was checking write-in responses to help identify race and Hispanic origin. In this procedure, write-in responses were computer checked against a large census database for these characteristics. If the system made a match, then it assigned a numeric code to the response. If no match was found, then clerks attempted to assign an appropriate race or Hispanic origin.

2. Methods

A separate document gives a high-level overview of the features of the census characteristic imputation system (Shores 2010). Census characteristic imputation contained two major components. These were the pre-edit and edit/allocation. The pre-edit cleaned and validated the data, and changed or set to blank data values in some cases. Once the pre-edit was completed, various edit and allocation processes filled in all remaining missing values.

The census system drew from hot decks to impute missing values when it could not use other methods of imputation. The hot decks were implemented by matrices whose cells were

categorized by attributes of persons in the household, the householder, or of the overall household, such as type of household or household composition.

Editing was a fundamental part of the census characteristic imputation system. The editing rules could alter the data to produce desired outcomes, such as those in effect for relationship, age, and sex, that would achieve "consistent" households. As an example, a parent was required to be at least 15 years older than his biological children. There were also many edit rules for race and Hispanic origin.

3. Limitations

Because the editing and imputation procedures of the census characteristic imputation system were different from the characteristic imputation procedures used for the 2000 A.C.E. Revision II, some caution may be called for in comparing the characteristic imputation results from the CCM and the A.C.E. Revision II P samples. This consideration does not apply to the E sample, however.

4. Results

Table 1 presents information about the effects of editing on the CCM data. It shows, for each characteristic, the number of cases for which values were changed because of the census edit and imputation system, and the percentage of the total number of records in the P sample that were changed through this editing. For this table, the number of records changed for a characteristic represents the number of times that a respondent-provided characteristic was changed, or edited, by the census editing rules.

Table 1. P-sample Person Records Changed by Edits

Characteristic	Records Changed	
	Number	Percent of P Sample
Relationship	4,949	1.3
Age	984	0.3
Sex	0	0.0
Race	828	0.2
Hispanic Origin	330	0.1
Tenure	0	0

The number of records changed through the census editing procedures was small relative to the total for all of the characteristics, though the amount of editing varied, from none for sex and tenure to 1.3% for relationship. For race and Hispanic origin, changes that were attributable to

the write-in responses are not counted as changes due to the editing process. Note that it is possible that the respondent-provided characteristics could be edited during the CCM Person Clerical Matching operations, prior to the CCM data being sent through the census edit and imputation system. These types of edits are also not counted as changes due to the editing process.

The percentages are based on the count of 392,711 records in the P sample. This included all person and household records captured from the sample interview. Some of these people were later coded as not being in the P sample.

Table 2 gives results for the write-in responses for race and Hispanic origin.

Table 2. Write-in Responses for the P sample

Characteristic	Records With a Write-in Response		Records Changed as a Result of Valid Write-in Response	
	Number	Percent of P sample	Number	Percent of P sample
Hispanic Origin	16,153	4.1	735	0.2
Race*	57,663	14.7	6,245	1.6

*At least one write-in race value

There were 735 records, 0.2% of the P sample, whose Hispanic origin changed because of a write-in value. The edits changed 734 from Hispanic to non-Hispanic, and one from non-Hispanic to Hispanic. There were 6,245 records, 1.6% of the P sample, for which race changed because of valid write-in responses (some write-in responses were not usable). We took as a change to race any change in the person's racial composition resulting from the write-in responses.

The system allowed people to write in up to four race responses, one each for American Indian and Alaska Native, Asian, Native Hawaiian and Pacific Islander, and Some Other Race. If a person selected one of these races in the interview, then he had the option of writing in a response for that race category. Sometimes when a person did this, the effect was to change the person's race composition. For example, there were cases for which someone selected Asian as his race, and then wrote "Afghan" for the Asian race write-in. The write-in changed the person's race composition from Asian to White and Asian. The general pattern was for the write-in responses to expand a person's racial composition to include more race categories. Over half of the changes were due to the Some Other Race write-in, indicating that many people who did not feel that the available race categories accurately represented their race chose Some Other Race, and then took advantage of the opportunity to write in a response.

To see the extent of changes for the individual race write-in responses, we examined records for which there was only a single write-in response. By so doing we could be sure that any observed changes to race occurred because of the entry in that specific race write-in response. Table 3

gives results for these outcomes. This table differs from Table 2, in which we counted all records with *at least one* race write-in response. Since the results in Table 3 exclude records with multiple write-ins, the total in Table 3 is less than the number given in the race row in Table 2.

Table 3. Records With Only One Race Write-in Response

Race	Records With One Race Write-in Response		Records Changed as a Result of a Valid Write-in Response	
	Number	Percent of P sample	Number	Percent of Write-ins
American Indian and Alaska Native	20,531	5.2	130	0.6
Asian	2,708	0.7	324	12.0
Native Hawaiian and Pacific Islander	1,008	0.3	241	23.9
Some Other Race	33,051	8.4	5,389	16.3
Total	57,298	14.6	6,084	10.6

Table 3 indicates that most changes to race that occurred because of a write-in response were made because of the entry for Some Other Race. Approximately 16% of the 33,051 write-in responses for Some Other Race resulted in a change to race, along with 24% of the write-in responses for Native Hawaiian and Pacific Islander, and 12% of those for Asian. The write-in for American Indian and Alaska Native produced relatively few changes.

Table 4 shows for each characteristic the percentage of persons in the P and E samples who had the characteristic imputed, as well as the percentage that had at least one of the characteristics imputed. In general, we took as our definition of imputed any record that was not considered to be "as reported." There was no imputation of relationship in 2000. The entries in Table 4 are unweighted.

Table 4. Imputation Rates in the 2010 and 2000 P and E Samples

Year	Sample	Total People	Percentage of people with imputed characteristic						Percent with at least one imputed characteristic
			Relationship	Age	Sex	Race	Hispanic Origin	Tenure	
2010	P sample	392,711	2.5	6.1	1.3	2.8	2.6	2.6	10.0
	E sample	383,537	2.2	5.5	1.6	4.2	4.8	3.2	15.4
2000	P sample	706,245	NA	2.5	1.7	1.4	2.4	1.9	5.5
	E sample	704,602	NA	3.1	0.3	3.5	3.6	3.8	11.2

2000 data source: Cantwell et al. (2001)

In the P sample, the percentages requiring imputation were higher for all characteristics in 2010 than in 2000, except for sex. The largest differences were for age and race, with the percentage requiring imputation for age over twice as high in 2010 as in 2000. The percentage in the P sample with at least one characteristic imputed was also higher in 2010 than in 2000. In the E sample, the percentages requiring imputation were higher in 2010 than in 2000 for all characteristics except tenure, including the percentage with at least one characteristic imputed.

References

Cantwell, P., McGrath, D., Nguyen, N., and Zelenak, M. F. (2001), "Accuracy and Coverage Evaluation: Missing Data Results," DSSD Census 2000 Procedures and Operations Memorandum Series B-7*.

Shores, R. (2010), "Census Imputation for 2010 – High Level Description," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-24.